

Exact Inference for Gaussian Process Regression in case of Big Data with the Cartesian Product Structure

Belyaev Mikhail

MIKHAIL.BELYAEV@DATADVANCE.NET

Institute for Information Transmission Problems, Bolshoy Karetny per. 19, Moscow, 127994, Russia

DATADVANCE, llc, Pokrovsky blvd. 3, Moscow, 109028, Russia

PreMoLab, MIPT, Institutsky per. 9, Dolgoprudny, 141700, Russia

Burnaev Evgeny

EVGENY.BURNAEV@DATADVANCE.NET

Institute for Information Transmission Problems, Bolshoy Karetny per. 19, Moscow, 127994, Russia

DATADVANCE, llc, Pokrovsky blvd. 3, Moscow, 109028, Russia

PreMoLab, MIPT, Institutsky per. 9, Dolgoprudny, 141700, Russia

Kapushev Yermek

ERMEK.KAPUSHEV@DATADVANCE.NET

Institute for Information Transmission Problems, Bolshoy Karetny per. 19, Moscow, 127994, Russia

DATADVANCE, llc, Pokrovsky blvd. 3, Moscow, 109028, Russia

Abstract

Approximation algorithms are widely used in many engineering problems. To obtain a data set for approximation a factorial design of experiments is often used. In such case the size of the data set can be very large. Therefore, one of the most popular algorithms for approximation — Gaussian Process regression — can be hardly applied due to its computational complexity. In this paper a new approach for Gaussian Process regression in case of factorial design of experiments is proposed. It allows to efficiently compute exact inference and handle large multidimensional data sets. The proposed algorithm provides fast and accurate approximation and also handles anisotropic data.

1. Introduction

Gaussian Processes (GP) have become a popular tool for regression which has lots of applications in engineering problems (Rasmussen & Williams, 2006). They combine flexibility of being able to approximate a wide range of smooth functions with simple structure of Bayesian inference and interpretable hyperparameters.

GP regression algorithm has $\mathcal{O}(N^3)$ time complexity and $\mathcal{O}(N^2)$ memory complexity, where N is a size of the training sample. For large training sets (ten thousands or more)

construction of GP regression becomes intractable problem on current hardware.

There is significant amount of research concerning sparse approximation of GP regression reducing run-time complexity to $\mathcal{O}(M^2N)$ for some $M \ll N$ (for example, M can be the size of a subsample used for sparse approximation). There are also methods based on Mixture of GPs and Bayesian Machine Committee. Overview of these methods can be found in (Rasmussen & Williams, 2006; Rasmussen & Ghahramani, 2001; Quiñero-Candela & Rasmussen, 2005).

Reduced run-time and memory complexity can be achieved not only by means of sparse approximations and Mixtures of GPs but also by taking into account a structure of a design of experiments. In engineering problems they often use a factorial design (Montgomery, 2006). That is there are several groups of variables, in each group variables take values from some discrete finite set. Such sets and corresponding groups of variables are called *factors*. Number of different values in a factor is called *factor size* and the values themselves are called *levels*. The Cartesian product of factors forms the training set. When the factorial design of experiments is used the size of the data set can be very large (as it grows exponentially with dimension of input variables).

There are several methods based on splines which consider

this special structure of the given data (Stone et al., 1997). A disadvantage of these methods is that they work only with one-dimensional factors and can't be applied to a more general case when factors are multidimensional. Another shortcoming is that such approaches don't have approximation accuracy evaluation procedure.

There are also several approaches for GP regression on a lattice based on block Toeplitz covariance matrix with Toeplitz blocks and circulant embedding (Zimmerman, 1989; Chan & Wood, 1997; Dietrich & Newsam, 1997). Such methods have $\mathcal{O}(N \log N)$ time complexity and $\mathcal{O}(N)$ memory complexity. They can be used only if all the factors are one-dimensional and after monotonic transformation of each factor (levels of each factor should be equally spaced). However, in the case of multidimensional factors the covariance matrix doesn't have the desired structure (it should be block Toeplitz with Toeplitz blocks) and therefore these approaches can't be generalized for such design of experiments.

There is another problem which we are likely to encounter. Factor sizes can vary significantly. Engineers usually use large factors sizes if corresponding input variables have big impact on function values otherwise the factors sizes are likely to be small, i.e. the factor sizes are often selected using knowledge from a subject domain (Rendall & Allen, 2008). For example, if it is known that dependency on some variable is quadratic then the size of this factor will be 3 as a larger size is redundant. Difference between factor sizes can lead to degeneracy of the GP model. We will refer to this property of data set as *anisotropy*.

In this paper we describe an approach that takes into account factorial nature of the design of experiments in general case of multidimensional factors and allows to efficiently calculate exact inference of GP regression. We also discuss how to choose initial values of parameters for the GP model and regularization in order to take into account possible anisotropy of the training data set.

1.1. Approximation problem

Let $f(\mathbf{x})$ be some unknown smooth function. The task is given a data set $\mathcal{D} = \{(\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}\}_{i=1}^N$ of N pairs of inputs \mathbf{x}_i and outputs y_i construct an approximation $\hat{f}(\mathbf{x})$ of the function $f(\mathbf{x})$ where outputs y_i are assumed to be noisy with additive i.i.d. Gaussian noise:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_{noise}^2). \quad (1)$$

1.2. Factorial design of experiments

In this paper a special case is considered when the design of experiments is factorial. Let us refer to sets of points $s_k = \{x_{i_k}^k \in X_k\}_{i_k=1}^{n_k}$, $X_k \subset \mathbb{R}^{d_k}$, $k = 1, K$ as *factors*. A set of points \mathbf{S} is referred to as a factorial design of experi-

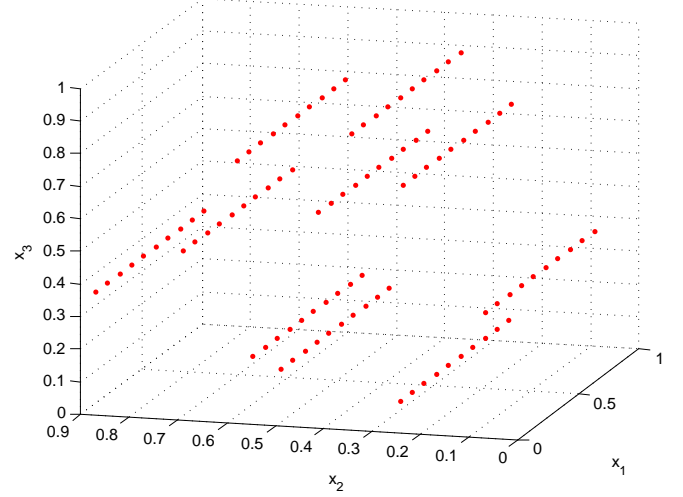


Figure 1. Example of a multidimensional factor. In the figure x_1 is a usual one-dimensional factor and (x_2, x_3) is a 2-dimensional factor.

ments if it is a Cartesian product of factors

$$\mathbf{S} = s_1 \times s_2 \times \dots \times s_K = \{[x_{i_1}^1, \dots, x_{i_K}^K], \{i_k = 1, \dots, n_k\}_{k=1}^K\}. \quad (2)$$

The elements of \mathbf{S} are vectors of a dimension $d = \sum_{i=1}^K d_i$ and the sample size is a product of sizes of all factors $N = \prod_{i=1}^K n_i$. If all the factors are one-dimensional \mathbf{S} is a full factorial design. But in a more general case factors are multidimensional (see example in Figure 1). Note that in this paper we do not consider categorical variables, factorial design is implemented across continuous real-valued features.

1.3. Gaussian Process Regression

GP regression is a Bayesian approach where a prior distribution over continuous functions is assumed to be a Gaussian Process, i.e.

$$\mathbf{f} | \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}_f), \quad (3)$$

where $\mathbf{f} = (f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_N))$ is a vector of outputs, $\mathbf{X} = (\mathbf{x}_1^T, \mathbf{x}_2^T, \dots, \mathbf{x}_N^T)^T$ is a matrix of inputs, $\boldsymbol{\mu} = (\mu(\mathbf{x}_1), \mu(\mathbf{x}_2), \dots, \mu(\mathbf{x}_N))$ is a mean vector for some function $\mu(\mathbf{x})$, $\mathbf{K}_f = \{k(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^N$ is a covariance matrix for some a priori selected covariance function k .

Without loss of generality we make the standard assumption of zero-mean data. Also assume that we are given observations with Gaussian noise

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_{noise}^2)$$

For a prediction of $f(\mathbf{x}_*)$ at new unseen data point \mathbf{x}_* the posterior mean conditioned on the observations $\mathbf{y} =$

(y_1, y_2, \dots, y_N) at training inputs \mathbf{X} is used

$$\hat{f}(\mathbf{x}_*) = \mathbf{k}(\mathbf{x}_*)^T \mathbf{K}_y^{-1} \mathbf{y}, \quad (4)$$

where $\mathbf{k}(\mathbf{x}_*) = (k(\mathbf{x}_*, \mathbf{x}_1), \dots, k(\mathbf{x}_*, \mathbf{x}_N))^T$, $\mathbf{K}_y = \mathbf{K}_f + \sigma_{noise}^2 \mathbf{I}$ and \mathbf{I} is an identity matrix. For approximation accuracy evaluation the posterior variance is used

$$\text{cov}(\hat{f}(\mathbf{x}_*)) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}(\mathbf{x}_*)^T \mathbf{K}_y^{-1} \mathbf{k}(\mathbf{x}_*). \quad (5)$$

Usually for GP regression a squared exponential covariance function is used

$$k(\mathbf{x}_p, \mathbf{x}_q) = \sigma_f^2 \exp \left(- \sum_{i=1}^d \theta_i^2 (\mathbf{x}_p^{(i)} - \mathbf{x}_q^{(i)})^2 \right). \quad (6)$$

Let us denote the vector of hyperparameters θ_i by $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$. To choose the hyperparameters of our model we consider the log likelihood

$$\begin{aligned} \log p(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}, \sigma_f, \sigma_{noise}) = & - \frac{1}{2} \mathbf{y}^T \mathbf{K}_y^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K}_y| - \\ & - \frac{N}{2} \log 2\pi \end{aligned} \quad (7)$$

and optimize it over the hyperparameters (Rasmussen & Williams, 2006). The runtime complexity of learning GP regression is $\mathcal{O}(N^3)$ as we need to calculate inverse of \mathbf{K}_y , its determinant and derivatives of the log likelihood.

2. Proposed approach

2.1. Tensor and related operations

For further discussions we will use tensor notation, so let's introduce definition of a tensor and some related operations.

A tensor \mathcal{Y} is a K -dimensional matrix of size $n_1 \times n_2 \times \dots \times n_K$ (Kolda & Bader, 2009):

$$\mathcal{Y} = \{y_{i_1, i_2, \dots, i_K}, \{i_k = 1, \dots, n_k\}_{k=1}^K\}. \quad (8)$$

By $\mathcal{Y}^{(j)}$ we will denote a matrix consisting of elements of the tensor \mathcal{Y} whose rows are $1 \times n_j$ slices of \mathcal{Y} with fixed indices $i_{j+1}, \dots, i_K, i_1, \dots, i_{j-1}$ and altering index $i_j = 1, \dots, n_j$. In case of 2-dimensional tensor it holds that $\mathcal{Y}^{(1)} = \mathcal{Y}^T$ and $\mathcal{Y}^{(2)} = \mathcal{Y}$.

Now let's introduce multiplication of a tensor by a matrix along the direction i . Let B be some matrix of size $n_i \times n'_i$. Then the product of the tensor \mathcal{Y} and the matrix B along the direction i is a tensor \mathcal{Z} of size $n_1 \times \dots \times n_{i-1} \times n'_i \times n_{i+1} \times \dots \times n_K$ such that $\mathcal{Z}^{(i)} = \mathcal{Y}^{(i)} B$. We will denote

this operation by $\mathcal{Z} = \mathcal{Y} \otimes_i B$. For a 2-dimensional tensor \mathcal{Y} multiplication along the first direction is a left multiplication by matrix $\mathcal{Y} \otimes_1 B = B^T \mathcal{Y}$, and along the second direction — is a right multiplication $\mathcal{Y} \otimes_2 B = \mathcal{Y} B$.

Multiplication of a tensor by a matrix along some direction is closely related to the Kronecker product. Let's consider an operation vec which for every multidimensional matrix \mathcal{Y} returns a vector containing all elements of \mathcal{Y} . An inner product of tensors \mathcal{Y} and \mathcal{Z} is the inner product of vectors $\text{vec}(\mathcal{Y})$ and $\text{vec}(\mathcal{Z})$

$$\langle \mathcal{Y}, \mathcal{Z} \rangle = \langle \text{vec}(\mathcal{Y}), \text{vec}(\mathcal{Z}) \rangle.$$

For every multidimensional matrix \mathcal{Y} of size $n_1 \times n_2 \times \dots \times n_K$ and $n_i \times p_i$ size matrices $B_i, i = 1, \dots, K$ the following identity holds (Kolda & Bader, 2009)

$$(B_1 \otimes B_2 \otimes \dots \otimes B_K) \text{vec}(\mathcal{Y}) = \text{vec}(\mathcal{Y} \otimes_1 B_1^T \otimes \dots \otimes_K B_K^T), \quad (9)$$

where symbol \otimes denotes the Kronecker product.

Let's compare a complexity of the right and the left hand sides of (9). For simplicity we assume that all the matrices B_i are quadratic of size $n_i \times n_i$ and $N = \prod n_i$. Then computation of the left hand side of (9) requires N^2 operations (of additions and multiplications) not taking into account complexity of the Kronecker product while the right hand side requires $N \sum_i n_i$ operations.

2.2. Computing inference

In this section an efficient way to compute inverse of a covariance matrix will be described as well as calculation of the log likelihood, its derivatives, the predictive mean and the covariance matrix using introduced tensor operations.

Covariance function (6) can be represented as a product of covariance functions each depending only on variables from one factor

$$k(\mathbf{x}_p, \mathbf{x}_q) = \prod_{i=1}^K k_i(x_p^i, x_q^i), \quad (10)$$

where $x_p^i, x_q^i \in \mathbb{R}^{d_i}$ belong to the same factor s_i . For the squared exponential function we have $k_i(x_p^i, x_q^i) = \sigma_{f,i}^2 \exp \left(- \sum_j^{d_i} \left(\theta_i^{(j)} \right)^2 \left(x_p^{(j),i} - x_q^{(j),i} \right)^2 \right)$, where $x_p^{(j),i}$ is a j -th component of x_p^i . Note that in general case covariance functions k_i are not necessarily squared exponential, they can be of different types for different factors. It allows to take into account special features of factors (knowledge from a subject domain) if they are known. In such case the function defined by (10) is still a valid covariance function being the product of separate covariance functions. From

now on we will denote by $\theta_i = (\theta_i^{(1)}, \dots, \theta_i^{(d_i)})$ the set of hyperparameters for covariance function of the i -th factor and let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$.

Such form of the covariance function and the factorial design of experiments allows us to represent the covariance matrix as the Kronecker product

$$\mathbf{K}_f = \bigotimes_{i=1}^K \mathbf{K}_i, \quad (11)$$

where \mathbf{K}_i is a covariance matrix defined by the k_i covariance function computed at points from the i -th factor s_i .

The Kronecker product of matrices can be efficiently inverted due to the following property

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$$

if A and B are invertible matrices. If A has size $n_a \times n_a$ and B has size $n_b \times n_b$ then the left side of the above equation requires $\mathcal{O}(n_a^3 n_b^3)$ operations while the right hand side requires only $\mathcal{O}(n_a^3 + n_b^3)$ operations and this is much less. However, we have to invert the matrix $\mathbf{K}_y = \mathbf{K}_f + \sigma_{noise}^2 \mathbf{I}$. For this we will use the Singular Value Decomposition (SVD)

$$\mathbf{K}_i = \mathbf{U}_i \mathbf{D}_i \mathbf{U}_i^T,$$

where \mathbf{U}_i is an orthogonal matrix of eigenvectors of matrix \mathbf{K}_i and \mathbf{D}_i is a diagonal matrix of eigenvalues. Using the properties of the Kronecker product and representing an identity matrix as $\mathbf{I}_{d_i} = \mathbf{U}_i \mathbf{U}_i^T$ we obtain

$$\mathbf{K}_y^{-1} = \left(\bigotimes_{i=1}^K \mathbf{U}_i \right) \left(\left[\bigotimes_{i=1}^K \mathbf{D}_i \right] + \sigma_{noise}^2 \mathbf{I} \right)^{-1} \left(\bigotimes_{i=1}^K \mathbf{U}_i^T \right). \quad (12)$$

Computing SVD for all \mathbf{K}_i requires $\mathcal{O}(\sum_k n_k^3)$ operations. Calculation of the Kronecker product in (12) has complexity $\mathcal{O}(N^2)$. So, this gives us overall complexity $\mathcal{O}(N^2)$ for calculation of expressions for the log likelihood, the predictive mean and the covariance. It is faster than the straightforward calculations, however it can be improved.

Equations (4), (5), (7) for GP regression do not require explicit inversion of \mathbf{K}_y . In each equation it is multiplied by vector \mathbf{y} (or \mathbf{k}_*). So, we will compute $\mathbf{K}_y^{-1} \mathbf{y}$ instead of explicitly inverting \mathbf{K}_y and then multiplying it by the vector \mathbf{y} .

Let \mathcal{Y} be a tensor containing values of the vector \mathbf{y} such that $\text{vec}(\mathcal{Y}) = \mathbf{y}$. Now using identities (9) and (12) we can

write $\mathbf{K}_y^{-1} \mathbf{y}$ as

$$\begin{aligned} \mathbf{K}_y^{-1} \mathbf{y} &= \left(\bigotimes_{i=1}^K \mathbf{U}_i \right) \left(\left[\bigotimes_{i=1}^K \mathbf{D}_i \right] + \sigma_{noise}^2 \mathbf{I} \right)^{-1} \times \\ &\quad \times \text{vec}(\mathcal{Y} \otimes_1 \mathbf{U}_1 \cdots \otimes_K \mathbf{U}_K) = \\ &= \text{vec} \left[\left((\mathcal{Y} \otimes_1 \mathbf{U}_1 \cdots \otimes_K \mathbf{U}_K) * \mathcal{D}^{-1} \right) \otimes_1 \right. \\ &\quad \left. \otimes_1 \mathbf{U}_1^T \cdots \otimes_K \mathbf{U}_K^T \right], \quad (13) \end{aligned}$$

where \mathcal{D} is a tensor constructed by transforming the diagonal of matrix $\left[\bigotimes_k \mathbf{D}_k \right] + \sigma_{noise}^2 \mathbf{I}$ into a tensor.

The elements of the tensor \mathcal{D} are eigenvalues of the matrix \mathbf{K}_y , therefore its determinant can be calculated as

$$|\mathbf{K}_y| = \prod_{i_1, \dots, i_K} \mathcal{D}_{i_1, \dots, i_K}. \quad (14)$$

Proposition 1. *The computational complexity of the log likelihood (7), where $\mathbf{K}_y^{-1} \mathbf{y}$ and $|\mathbf{K}_y|$ are calculated using (13) and (14), is*

$$\mathcal{O} \left(\sum_{i=1}^K n_i^3 + N \sum_{i=1}^K n_i \right). \quad (15)$$

Proof. Let's calculate the complexity of computing $\mathbf{K}_y^{-1} \mathbf{y}$ using (13). Computation of the matrices \mathbf{U}_i and \mathbf{D}_i requires $\mathcal{O}(\sum_i n_i^3)$ operations. Multiplication of the tensor \mathcal{Y} by the matrices \mathbf{U}_i requires $\mathcal{O}(N \sum_i n_i)$ operations. Further, component-wise product of the obtained tensor and the tensor \mathcal{D}^{-1} requires $\mathcal{O}(N)$ operations. And complexity of multiplication of the result by the matrices \mathbf{U}_i is again $\mathcal{O}(N \sum_i n_i)$. The determinant, calculated by equation (14), requires $\mathcal{O}(N)$ operations. Thus, the overall complexity of computing (13) is $\mathcal{O}(\sum_{i=1}^K n_i^3 + N \sum_{i=1}^K n_i)$. \square

For more illustrative estimate of the computational complexity suppose that $n_i \ll N$ (number of factors is large and their sizes are close). In this case it holds that $\mathcal{O}(N \sum_i n_i) = \mathcal{O}(N^{1+\frac{1}{K}})$ and this is much less than $\mathcal{O}(N^3)$.

To optimize the log likelihood over the hyperparameters we use a gradient based method. The derivatives of the log likelihood with respect to the hyperparameters take the form

$$\begin{aligned} \frac{\partial}{\partial \theta} (\log p(\mathbf{y} | \mathbf{X}, \sigma_f, \sigma_{noise})) &= \\ &= -\frac{1}{2} \text{Tr}(\mathbf{K}_y^{-1} \mathbf{K}') + \frac{1}{2} \mathbf{y}^T \mathbf{K}_y^{-1} \mathbf{K}' \mathbf{K}_y^{-1} \mathbf{y}, \quad (16) \end{aligned}$$

where θ is one of the hyperparameters of covariance function (component of θ_i , σ_{noise} or $\sigma_{f,i}$, $i = 1, \dots, d$) and

$\mathbf{K}' = \frac{\partial \mathbf{K}}{\partial \theta}$. \mathbf{K}' is also the Kronecker product

$$\mathbf{K}' = \mathbf{K}_1 \otimes \cdots \otimes \mathbf{K}_{i-1} \otimes \frac{\partial \mathbf{K}_i}{\partial \theta} \otimes \mathbf{K}_{i+1} \otimes \cdots \otimes \mathbf{K}_K,$$

where θ is a parameter of the i -th covariance function. Denoting by \mathcal{A} a tensor such that $\text{vec}(\mathcal{A}) = \mathbf{K}_y^{-1} \mathbf{y}$ the second term in equation (16) can be efficiently computed using the same technique as in (13):

$$\begin{aligned} \frac{1}{2} \mathbf{y}^T \mathbf{K}_y^{-1} \mathbf{K}' \mathbf{K}_y^{-1} \mathbf{y} = \\ \left\langle \mathcal{A}, \mathcal{A} \otimes_1 \mathbf{K}_1^T \otimes_2 \cdots \otimes_{i-1} \mathbf{K}_{i-1}^T \otimes_i \right. \\ \left. \frac{\partial \mathbf{K}_i^T}{\partial \theta} \otimes_{i+1} \mathbf{K}_{i+1}^T \otimes_{i+2} \cdots \otimes_K \mathbf{K}_K^T \right\rangle. \end{aligned} \quad (17)$$

The complexity of calculating this term of derivative is the same as the complexity of equation (13).

Now let's compute the first term

$$\begin{aligned} \text{Tr}(\mathbf{K}_y^{-1} \mathbf{K}') &= \text{Tr} \left(\left(\bigotimes_{i=1}^K \mathbf{U}_i \right) \mathbf{D}^{-1} \left(\bigotimes_{i=1}^K \mathbf{U}_i^T \right) \mathbf{K}' \right) = \\ &= \text{Tr} \left(\mathbf{D}^{-1} \left(\bigotimes_{i=1}^K \mathbf{U}_i^T \mathbf{K}'_i \mathbf{U}_i \right) \right) = \\ &= \left\langle \text{diag}(\mathbf{D}^{-1}), \text{diag} \left(\bigotimes_{i=1}^K \mathbf{U}_i \mathbf{K}'_i \mathbf{U}_i \right) \right\rangle = \\ &= \left\langle \text{diag}(\mathbf{D}^{-1}), \bigotimes_{i=1}^K \text{diag}(\mathbf{U}_i \mathbf{K}'_i \mathbf{U}_i) \right\rangle, \end{aligned} \quad (18)$$

where $\text{diag}(\mathbf{A})$ is a vector of diagonal elements of a matrix \mathbf{A} , $\mathbf{D} = \bigotimes_i \mathbf{D}_i + \sigma_{\text{noise}}^2 \mathbf{I}$.

The computational complexity of this derivative term is the same as the computational complexity of equation (17).

Thus, we obtain

Proposition 2. *The computational complexity of calculating derivatives of the log likelihood is $\mathcal{O} \left(\sum_{i=1}^K n_i^3 + N \sum_{i=1}^K n_i \right)$.*

Table 1 contains training times for original GP and proposed GP regression for different sample sizes. The experiments were conducted on a PC with Intel i7 2.8 GHz processor and 4 GB RAM. For original GP we used GPML Matlab code (Rasmussen & Nickisch, 2010). We also adopted GPML code to use tensor operations. The results illustrate that the proposed approach is much more faster than the original GP and allows to make approximations using extremely large data sets.

Table 1. Runtime (in seconds) of tensored GP and original GP algorithms.

	ORIGINAL GP	TENSORED GP
64	0.8	0.16
160	2.69	0.16
432	14.31	0.74
1000	120.38	1.02
2000	970.21	1.11
10240	—	33.18
64000	—	74.9
160000	—	175.15
400000	—	480.14

2.3. Anisotropy

In this section we will consider an anisotropy problem. As it was mentioned in an engineering practice factorial designs are often anisotropic, i.e. sizes of factors differ significantly. It is a common case for the GP regression to become degenerate in such situation. Suppose that the given design of experiments consists of two one-dimensional factors with sizes n_1 and n_2 . Assume that $n_1 \ll n_2$. Then one could expect the length-scale for the first factor to be much greater than the length-scale for the second factor (or $\theta_1 \ll \theta_2$). However, in practice we often observe the opposite $\theta_1 \gg \theta_2$. This happens because the optimization algorithm stacks in a local maximum during maximization over the hyperparameters as the objective function (the log likelihood) is non-convex with lots of local maxima. We get an undesired effect of degeneracy: in the region without training points the approximation is constant and it has sharp peaks at training points. This situation is illustrated in Figure 3 (compare with the true function in Figure 4).

Let us denote length-scales as $l_k^{(i)} = [\theta_k^{(i)}]^{-1}$. To incorporate our prior knowledge about factor sizes into regression model we introduce prior distribution on the hyperparameters θ :

$$\frac{\theta_k^{(i)} - a_k^{(i)}}{b_k^{(i)} - a_k^{(i)}} \sim \mathcal{Be}(\alpha, \beta), \quad \{i = 1, \dots, d_k\}_{k=1}^K, \quad (19)$$

i.e. prior on hyperparameter $\theta_k^{(i)}$ is a beta distribution with parameters α and β scaled to some interval $[a_k^{(i)}, b_k^{(i)}]$.

The log likelihood then has the form

$$\begin{aligned} \log p(\mathbf{y} | \mathbf{X}, \boldsymbol{\theta}, \sigma_f, \sigma_{noise}) = & -\frac{1}{2} \mathbf{y}^T \mathbf{K}_y^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K}_y| - \\ & -\frac{N}{2} \log 2\pi + \sum_{k,i} \left((\alpha - 1) \log \left(\frac{\theta_k^{(i)} - a_k^{(i)}}{b_k^{(i)} - a_k^{(i)}} \right) + \right. \\ & \left. + (\beta - 1) \log \left(1 - \frac{\theta_k^{(i)} - a_k^{(i)}}{b_k^{(i)} - a_k^{(i)}} \right) \right) - d \log(B(\alpha, \beta)), \end{aligned} \quad (20)$$

where $B(\alpha, \beta)$ is a beta function.

Numerous references use gamma distribution as a prior, e.g. (Neal, 1997). Preliminary experiments showed that GP regression models with gamma prior often degenerates. So, in this work we use prior with compact support. By introducing such prior we restrict parameters $\theta_k^{(i)}$ to belong to some interval $[a_k^{(i)}, b_k^{(i)}]$ (or length-scales $l_k^{(i)}$ to belong to the interval $[(b_k^{(i)})^{-1}, (a_k^{(i)})^{-1}]$). This choice of prior allows to prohibit too small and too large length-scales excluding possibility to degenerate (if intervals of allowed length-scales are chosen properly).

It seems reasonable that for an approximation to fit the training points the length-scale is not needed to be much less than the distance between points. That's why we choose the lower bound for the length-scale $l_k^{(i)}$ to be $c_k * \min_{x, y \in s_k, x^{(i)} \neq y^{(i)}} \|x^{(i)} - y^{(i)}\|$ and the upper bound for the length-scale to be $C_k * \max_{x, y \in s_k} \|x^{(i)} - y^{(i)}\|$. The value c_k should be close to 1. If it is chosen too small we are taking risks to overfit the data by allowing small length-scales. If c_k is too large we are going to underfit the data by allowing only large length-scales and forbidding small ones. Constants C_k must be much greater than c_k to permit large length-scales and preserve flexibility. In this work we used $c_k = 0.5$ and $C_k = 100$. Such values of c_k and C_k worked rather good in our test cases.

Parameters of beta distribution was set to $\alpha = \beta = 2$ to get symmetrical probability distribution function (see Figure 2). We don't know a priori large or small should be the values of GP parameters, so prior distribution should impose nearly the same penalties for the intermediate values of the parameters. As it can be seen from Figure 2 the chosen distribution does exactly what is needed.

Figure 5 illustrates approximation of the GP regression with introduced prior distribution (and initialization described in Section 2.4). The hyperparameters were chosen such that the approximation is non-degenerate.

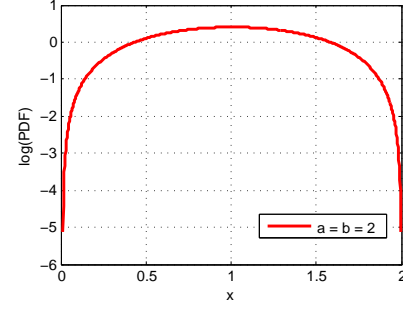


Figure 2. Logarithm of Beta distribution probability density function, rescaled to $[0.01, 2]$ interval, with parameters $\alpha = \beta = 2$.

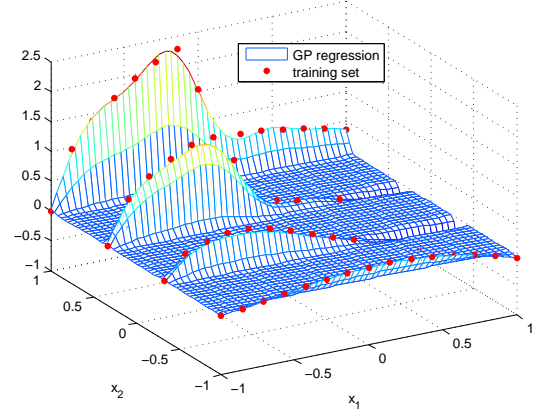


Figure 3. Degeneracy of the GP model in case of anisotropic data set. The length-scales for this model are $l_1 = 0.286$, $l_2 = 0.033$ whereas factor sizes are $n_1 = 15$, $n_2 = 4$.

2.4. Initialization

It is also important to choose reasonable initial values of hyperparameters in order to converge to a good solution during parameters optimization. The kernel-widths for different factors should have different scales because corresponding factor sizes have different number of levels. So it seems reasonable to use average distance between points in a factor as an initial value

$$\theta_k^{(i)} = \left[\frac{1}{n_k} \left(\max_{x \in s_k} (x^{(i)}) - \min_{x \in s_k} (x^{(i)}) \right) \right]^{-1}. \quad (21)$$

3. Experimental results

Proposed algorithm was tested on a set of test functions (Evolutionary computation pages — the function testbed; System optimization — testproblems). The functions have different input dimensions from 2 to 6 and the sample sizes N varied from 100 to about 200000. For each function several factorial anisotropic training sets were generated. We will test the following algorithms: GP with tensor computations (tensorGP), GP with tensor computations and prior

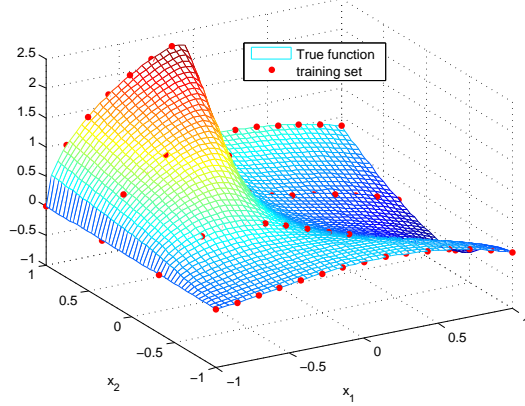


Figure 4. True function.

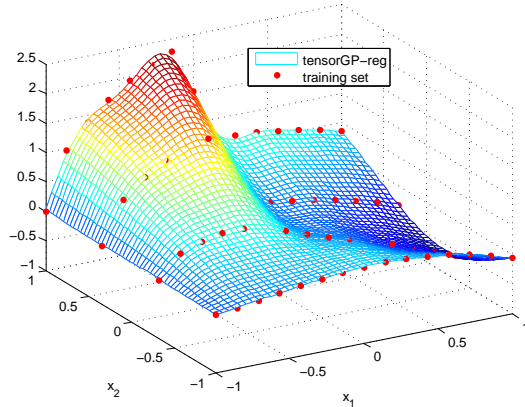


Figure 5. The GP regression with proposed prior distribution and initialization.

distribution (tensorGP-reg), the sparse pseudo-point input GP (FITC) (Snelson & Ghahramani, 2005). For FITC method we used GPML Matlab code (Rasmussen & Nickisch, 2010). Number of inducing points of FITC algorithm varied from $M = 500$ for small samples (up to 5000 points) to $M = 70$ for large samples (about 10^5 points) in order to obtain approximation in reasonable time (complexity of FITC algorithm is $\mathcal{O}(M^2N)$). For tensorGP and tensorGP-reg we adopted GPML code to use tensor operations, proposed prior distribution and initialization.

To assess quality of approximation a mean squared error was used

$$\text{MSE} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} (\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i))^2, \quad (22)$$

where $N_{\text{test}} = 50000$ is a size of test set. The test sets were generated randomly.

A large set of problems (number of problems is about 40) was used to test the algorithms. Usual “test problem vs. MSE” plot will be confusing for large set of problems as it will look like noisy graph. To see a picture of the overall performance of algorithms on such test problems set we use Dolan-Moré curves (Dolan & Moré, 2002). The idea of Dolan-Moré curves is as follows. Let $t_{p,a}$ be an error of an a -th algorithm on a p -th problem and $r_{p,a}$ be a performance ratio

$$r_{p,a} = \frac{t_{p,a}}{\min_s(t_{p,s})}.$$

Then Dolan-Moré curve is a graph of $\rho_a(\tau)$ function where

$$\rho_a(\tau) = \frac{1}{n_p} \text{size}\{p : r_{p,a} \leq \tau\},$$

which can be thought of as a probability for the a -th algorithm to have performance ratio within factor $\tau \in \mathbb{R}_+$. The higher the curve $\rho_a(\tau)$ is located the better works the corresponding algorithm. $\rho_a(1)$ is a number of problems on which the a -th algorithm showed the best performance.

As expected tensorGP performs better than FITC as it uses all the information containing in the training sample. Introduced prior distribution is more suited for anisotropic data that’s why GP with such prior (tensorGP-reg) performs even better (see Figure 6).

To compare run-time performances of the algorithms we plotted Dolan-Moré curves where instead of approximation error the training time was used (see Figure 7). Here we see that tensorGP and tensorGP-reg outperform FITC algorithm.

3.1. Rotating disc problem

In this section we will consider a real world problem of rotating disc shape design. Such kind of problems often

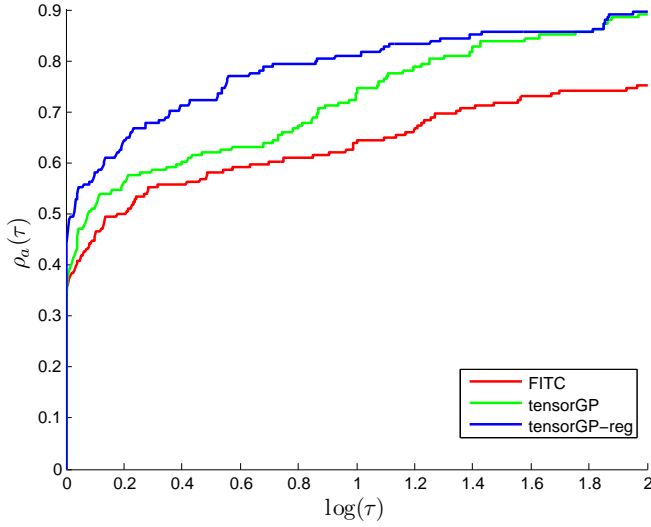


Figure 6. Approximations accuracies comparison. Dolan-Moré curves for tensorGP, tensorGP-reg and FITC algorithms in logarithmic scale. The higher lies the curve the better performs the corresponding algorithm.

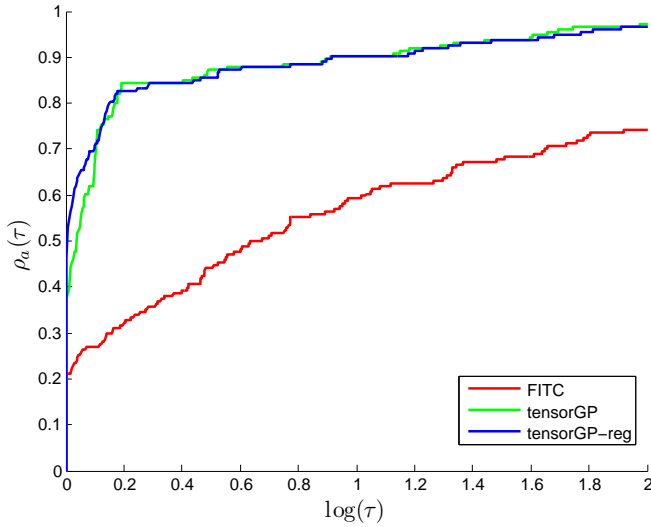


Figure 7. Run-times comparison. Dolan-Moré curves for tensorGP, tensorGP-reg and FITC algorithms in logarithmic scale. The higher lies the curve the better performs the corresponding algorithm.

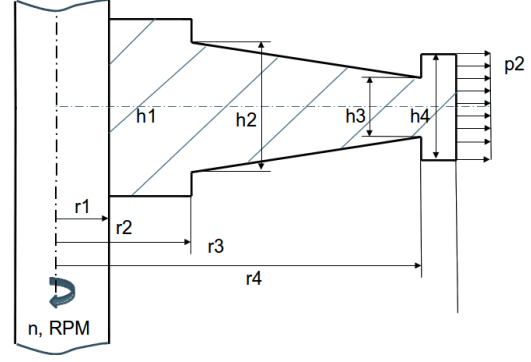


Figure 8. Rotating disc parametrization.

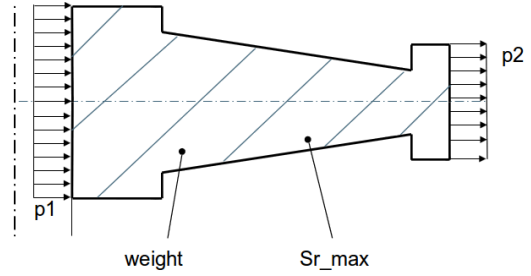


Figure 9. Rotating disc objectives.

arises during aircraft engine design and in turbomachinery (Armand, 1995).

In this problem a disc of an impeller is considered. It is rotated around the shaft. The geometrical shape of the disc considered here is parameterized by 6 variables $\mathbf{x} = (h_1, h_2, h_3, h_4, r_2, r_3)$ (r_1 and r_4 are fixed), see Figures 8 and 9. The task of an engineer is to find such geometrical shape of the disc that minimizes disc's weight and contact pressure p_1 between the disc and the shaft while constraining the maximum radial stress Sr_{max} to be less than some threshold. The physical model of a rotating disc is described in (Armand, 1995) and it was adopted to the disc shape presented in Figures 8, 9 in order to calculate the contact pressure p_1 and the maximum radial stress Sr_{max} .

It is a common practice to build approximations of objective functions in order to analyze them and perform optimization (Forrester et al., 2008). So, we applied developed in this work tensorGP-reg algorithm and FITC to this problem. The design of experiments was full factorial, number of points in each dimension was $[1, 8, 8, 3, 15, 5]$, i.e. x_1 was fixed. Number of points in factors differ significantly and the generated data set is anisotropic. The overall number of points in the training sample was 14 400.

Figures 10 and 11 depict 2D slices of contact pressure approximations along x_5, x_6 variables. As you can see FITC model degenerates while tensorGP-reg provides smooth

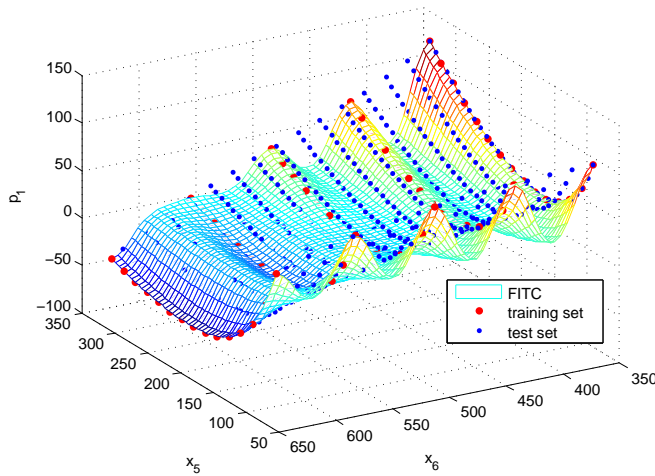


Figure 10. 2D slice along x_5 and x_6 variables (other variables are fixed) of FITC approximation with 500 inducing inputs. It can be seen that the approximation degenerates.

and accurate approximation.

4. Conclusion

Gaussian Processes are often used for building approximations for small data sets. However, knowledge about the structure of the given data set can contain important information which allows us to efficiently compute exact inference even for large data sets.

Introduced prior distribution combined with reasonable initialization has proven to be an efficient way to struggle degeneracy in case of anisotropic data.

Algorithm proposed in this paper takes into account the special factorial structure of the data set and is able to handle large multidimensional samples preserving power and flexibility of GP regression. Our approach has been successfully applied to toy and real problems including the rotating disc shape design.

References

- Armand, S. C. *Structural Optimization Methodology for Rotating Disks of Aircraft Engines*. NASA technical memorandum. National Aeronautics and Space Administration, Office of Management, Scientific and Technical Information Program, 1995.
- Chan, Grace and Wood, Andrew T.A. Algorithm as 312: An algorithm for simulating stationary gaussian random fields. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(1):171–181, 1997. ISSN 1467-9876.
- Dietrich, C. R. and Newsam, G. N. Fast and exact simu-

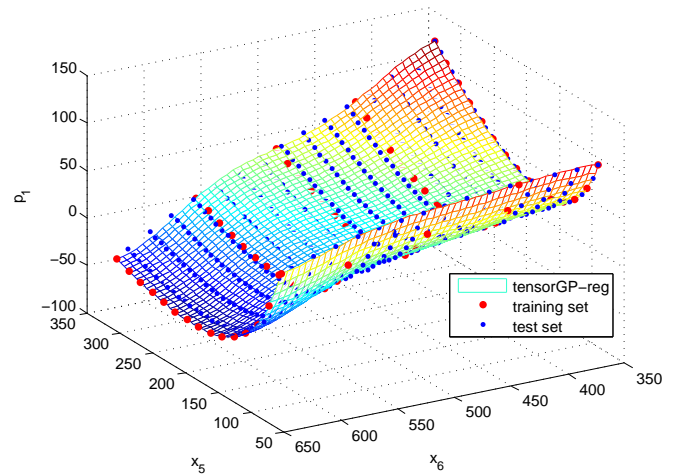


Figure 11. 2D slice along x_5 and x_6 variables (other variables are fixed) of tensorGP-reg approximation. It can be seen that tensorGP-reg provides accurate approximation.

lation of stationary gaussian processes through circulant embedding of the covariance matrix. *SIAM J. Sci. Comput.*, 18(4):1088–1107, July 1997. ISSN 1064-8275.

Dolan, Elizabeth D. and Moré, Jorge J. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91(2):201–213, January 2002.

Evolutionary computation pages — the function testbed. *Laappeenranta University of Technology*. URL <http://www.it.lut.fi/ip/evo/functions/functions.html>.

Forrester, Alexander I. J., Sobester, Andras, and Keane, Andy J. *Engineering Design via Surrogate Modelling - A Practical Guide*. J. Wiley, 2008.

Kolda, Tamara G. and Bader, Brett W. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

Montgomery, Douglas C. *Design and Analysis of Experiments*. John Wiley & Sons, 2006. ISBN 0470088109.

Neal, Radford M. Monte carlo implementation of gaussian process models for bayesian regression and classification. *arXiv preprint physics/9701026*, 1997.

Quiñonero-Candela, Joaquin and Rasmussen, Carl Edward. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

Rasmussen, Carl E. and Williams, Christopher. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

- Rasmussen, Carl Edward and Ghahramani, Zoubin. Infinite mixtures of gaussian process experts. In *In Advances in Neural Information Processing Systems 14*, pp. 881–888. MIT Press, 2001.
- Rasmussen, Carl Edward and Nickisch, Hannes. Gaussian processes for machine learning (gpml) toolbox. *J. Mach. Learn. Res.*, 11:3011–3015, December 2010. ISSN 1532-4435.
- Rendall, T.C.S. and Allen, C.B. Multi-dimensional aircraft surface pressure interpolation using radial basis functions. *Proc. IMechE Part G: Aerospace Engineering*, 222:483 – 495, 2008. Publisher: IMechE.
- Snelson, Edward and Ghahramani, Zoubin. Sparse gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems 18*, pp. 1257–1264. MIT press, 2005.
- Stone, Charles J., Hansen, Mark, Kooperberg, Charles, and Truong, Young K. Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist.*, 25: 1371–1470, 1997.
- System optimization — testproblems. *Swiss International Institute of Technology*. URL <http://www.tik.ee.ethz.ch/sop/download/supplementary/testproblems/>.
- Zimmerman, DaleL. Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. *Mathematical Geology*, 21(7):655–672, 1989. ISSN 0882-8121.